Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 235 – 240

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Data Mining Applications in E-Government Information Security

Tongwei yuan*, Peng Chen

Software Engineering Center Chinese Academy of Sciences, 4S 4th St. ZhongGuan Cun P.O.Box2717, Beijing 100190,China

Abstract

This paper proposes a current more mature Association Analysis Model after it gives an overview of data mining method, and conducts a formal description. Various industries in the current increasing emphasis on information security, we give an Information Security Risk Assessment for a large-scale e-government by using Association Analysis Model. This method provides valuable reference data for enterprise information security assessment and decision analysis.

Keyword: Data mining; E-government; information security

1. Introduction

In twenty-first Century, information technology has rapidly permeated into every field of human society. Form the Large national or international fields to the small families or individuals, more and more people use of information technology to provide convenient, fast and efficient work and business. The expanding demands of All walks of life prompted the investment and the scale of informatization construction from the underlying physical layer to the top of the application system constantly strengthen intensify. In the imperceptibly we have entered into a new era, the information age.

But the information system development is a "double-edged sword". It created enormous benefits for the mankind, at the same time, the issue of information security was to bring a great loss and inconvenience. The virus, hackers rampant, the leakage of secrets, system failure, interruption of service

* Corresponding author. Tel.: +86-010-6253-7859.
E-mail address: twyuan@sec.ac.cn.

doi:10.1016/j.proeng.2011.12.700

and various computer crimes emerged in an endless stream. With the human's high dependence on informatization, loss caused by information security is becoming more and more striking. According to the United States of FBI survey, economic losses caused by the network security are more than $170 billion in USA per year. 75% companies reported that the financial losses were caused by results of computer system security problems[1]. From CN CERT and the China Internet Network Information Center's annual reports, in the first half of 2010, CN CERT had received 4780 network security incident reports, increase of 105%. In the past year, the service fee expenditure for processing safety events totalling up to 153 billion yuan[2].

Therefore, the security problem of the information system is more and more highly attracted the attention of various government sector. They are taking various means of supervision, urging agencies at all levels to improve information security awareness and information security measures, avoiding the risk of information security. At present, a lot of supervision mechanisms collect a lot of information security data through a variety of ways. We give out an correlation analysis on the information security of data based on the basing on data mining technology, in order to find out a good method of information security assessment for the governments and enterprises, and assisting the information security supervision departments to make effective decisions.

2. Data mining method overview

2.1. K-means algorithm

K-means algorithm (Lloyd, 1982) is a simple and effective statistical clustering technology, it gives specific classes number K, and put N objects into the K classes, to make the maximum similarity within objects in any class, and to make the minimum similarity among class. The algorithm first need to make an initial judgment, That is to select the initial class number and the initial cluster center, and then, each sample is placed in the similar class. Similarity measure can be defined in many different ways. The most commonly used similarity metric is simple Euclidean distance. After all samples are placed into the appropriate class, the class center was update through the calculation of each new class of average. The process is repeated until a certain iteration of generating class center no longer change.

2.2. Decision tree

The decision tree algorithm was first proposed as ID3 algorithm by Quinlan, Later there reappeared many kinds of decision tree algorithm such as ID4, ID5, C4.5, CART, CLOUDS, PUBLIC, SLIQ, RAINFOREST, SPRINT, ScalParC and so on. It is a common structure to supervise learning, Mainly used for data classification. First, we should select portion of the samples to create a decision tree from the training set, and the remainder of the training samples are used to inspect the accuracy of the tree established. If the decision tree can correctly classify the remaining samples, the process will be end. If some sample's classification is error, this sample is added to the training set and create a new tree. In this way, we design a tree which can classify all training samples correctly.

2.3. Artificial neural network

Artificial neural network was born in 1950, and Rosenblatt put the single-layer perception application in pattern classification. Its principle is the human brain thinking system's simple structure simulation. It's a multilayered network that is made up of a number of neuronal connections, and can imitate the human brain function of neuron. It is also the adaptive function estimator that does not rely on the model.

It does not need any model to realize arbitrary function relation. Its advantage is capable of parallel processing, and has the learning ability, adaptability and strong fault tolerant ability.

3. The formal description for correlation analysis model

Correlation analysis model is adapt to find out the meaningful connection hiding in the large data sets. The connection found can be represented by association rules or frequent itemsets form. We need to deal with two critical issues in the data correlation analysis. First, finding in computing mode from the large object data concentration may be costly, second, some models found may be false, because they may have occurred by chance.

3.1. Data's two elements expression

The hypothesis that certain things and the Item set included in these things in the set of relations as shown in Table 1:

Table 1 . Data's two elements 0/1 expression

TID	i1	i2	i3	i4	i5	i6
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

If I={i1,i2,…,id} is the collection of all items, and T={t1,t2,…,tn} is the collection of all things. Each thing of ti contains item set is the subset of I, that is ti being contained in I. Item set's an important property is the support count, that is the number of things containing specific item sets, With σ (X) expressed as:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}| \quad (1)$$

For example, table 1 shows the data concentration, item set $\{t_2,t_3,t_4\}$ support count is 2, because only these 2 things at the same time contain these 3 items. [5]

3.2. The support degree and confidence degree

Support degree is an important measure, because the rules which support degree is very low may sometimes occur. From the data analysis point of view, most of low support degree of rules will be meaningless, because it was more important to improve the research and find out countermeasures to focus on more important things than on the things which does not often happen. Therefore, when we carry out data analysis, we can use the support threshold interval to delete those meaningless rules. In addition, support degree also has a desired properties, can be used for association rules effectively discover.

The confidence degree measures the reasoning reliability by rules or mathematical model. For a given rule X→Y, the higher the confidence degree, the more likely that Y will appear in the things which contains X. The confidence degree can also estimate the Y defined in X under the conditional probability.

If the association rules are expressed as shaped like X→Y implication expression, where X and Y are disjoint sets, namely X∩Y=ø. The strength of association rule can be measured by support degree and

confidence degree. Support degree determines the rules that can be used for a given data set frequency, while the confidence degree determines the frequent degree that Y in X contains things appeared. Support degree (s) and confidence degree (c) these two metric representation just as below:

$$s(X \to Y) = \frac{\sigma\ (X \cup Y)}{N} \tag{2}$$

$$c(X \to Y) = \frac{\sigma\ (X \cup Y)}{\sigma\ (X)} \tag{3}$$

Support degree measures the importance (or range) of association rule, confidence degree measures the accuracy of association rules. At the same time satisfying the minimum support threshold (min-support) and minimum confidence threshold (min-confidence) rules called the strong rule. The problem of association rule mining is the strong rule, that is satisfied for min-support and min-confidence at the same time when mining in transaction database. [5]

3.3. Algorithm optimization

Mining association rules in a primitive method is to calculate for each possible rules support degree and confidence degree. But this method is costly, and step back, because of the number of rules extracted from the data set up to index level. For example, the total number of rules extracted from a data set containing d is, $R=3^d-2^{d+1}+1$. If there are 6 items of data set, it need to calculate the 602 rules of the support degree and confidence degree.

The optimization strategy of mining algorithm of association rules is to decompose the association rules mining task into two processes. The first is the frequent itemsets generation process, that is the process finding to meet the minimum support threshold. The second process is the generation rules, that is to extract all the high confidence rules from frequent itemsets found from the previous step. There are many optimization algorithm in frequent itemsets generation process, such as the transcendental principle, apriori algorithm, candidate itemsets generation and pruning algorithm.

4. Association analysis application in the analysis of Information Security

In this part, we give an IT environment of the government with hundreds of branches as a case. We give an information security risk assessment analysis on years of Information Security Survey and information security events and other research data according to the proposed model relational analysis, and give a step by step sample. This paper describes the data up to the end of 2010 census data completed, and streamlines a part of the data to be example.

Exclusion of different industry, the information security census data is up to thousands items, users wanted to investigate the relationship amount some projects according to some empirical, and analyzed some valuable information related with information security from the artificial.

4.1. Data example and transformation to two elements table

Although the user's information security census data is up to tens of thousands of items, but we did not think it was necessary to make correlation analysis on thousands of items at the same time. Users according to their experiences can first find out the items which are inner linked to each other, and then make a correlation analysis through the model applied. Users gave a set of information security data sets as shown below, We use a certain algorithm to process data, because the data must be kept confidential.

Table 2 Information security data extraction of all organization

TID	Company scale	last year IT invest prop	IT person Proportion	Information security score	security event	……
1	2.5	3%	7%	85	happen	……
2	10.7	1.5%	3%	73	happen	……
3	15.8	8%	10%	90	happen	……
4	4.2	5.6%	5%	60	Not happen	……
5	……	……	……	……	……	……

Company scale: Company total assets.

Last year IT investment Proportion: The same period last year the company all of the IT system hardware and software company total assets divided by the total value of investment.

IT person Proportion: the total number of IT person divided by the company formal permanent staff.

Information security score: The current on the company's information security aspects of total score.

security event: If the company has happened information security events or not in selected period.

Because the data in Table 2 don't meet the correlation analysis model needed two elements representation, we should first convert each item to multiple items. Its purpose is to adapt the existing association rules algorithms of data mining. Only to company scale and Last year IT investment Proportion for example can be transformed into below table:

Table 3 Item split and two elements expression

TID	scale <5	scale 5≥ <10	scale 10≤	invest Prop <3	Invest Prop 3≥ <6	Invest Prop 6≤	……
1	1	0	0	0	1	0	……
2	0	0	1	1	0	0	……
3	0	0	1	0	0	1	……
4	1	0	0	0	1	0	……

4.2. result analysis

In fact, making inferences through association rules does not necessarily contain the causal relationship, it just means the rules before the piece and the rear piece clearly appearing at the same time. We can give an artificial analysis and judgment for it, to find out some rules and reasons hidden in them. On the other hand, due to infer causality need knowledge of the causes and results attributes about data, sometimes we found that there is a correlation between the real cause of items by long-term observation.

We and the information security supervision department separately analyzed the 2009 and 2010 data of hundreds of units, and found that there implied a rule in these data:

R1:company scale∈[10,∞]→security event=happen (s=10%，c=75%)

We found the reasons through our analysis, one is the larger organization scale the more complex the informatization construction, the information security event happened more probably. Another more important reason is the larger organization scale more standardized on the unit management, they reported the information security event instantly once happened. Therefore, we strengthened the information security event report management system, and we increased strength of rewards and punishments, then according to the new stage data for further analysis.

Using correlation analysis model we undertook associated analysis between the number of the information security events and IT person Proportion, found two more interesting rules as below:

R2: 2009 year IT person Proportion∈[8%,∞]→security event∈[2,∞]=happen (s=50%，c=80%)

R3: 2010 year IT person Proportion∈[8%,∞]→security event∈[2,∞]=not happen (s=50%，c=80%)

We found the reason basing on the analysis of various factors. Although all the organizations were required to have complete requirement and system, but there had no measures. After 2008, they Increased strict supervision of the information safety, assessment of the strength, and formulate corresponding incentive policy. It seems obvious effect.

4.3. building the structure to support for correlation analysis model

In order to let the user can freely define correlation analysis elements, we provide a relational model database, to support the user customizing: item set, things condition, minimum support threshold, the minimum threshold. The table structure design is simplified as Fig. 1:

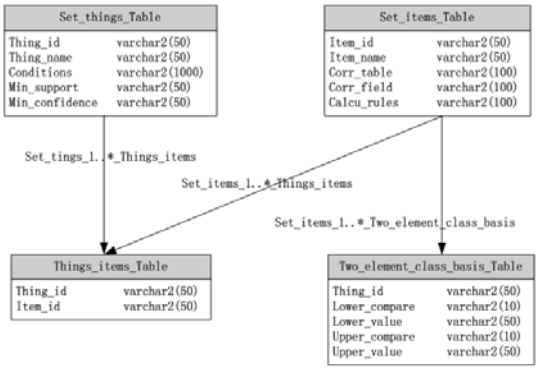

Fig. 1. Table structure design

The user can freely set in which kind of things made correlation analysis through a given configuration interface. First user should set things and items respectively, and then set filter conditions rules, min-support, and min-confidence when set things. Users should specify the item set value's calculation rules, when they set the items sets. At the same time, they should set two elements classification basis for each two elements item. Finally, users set the relationship between tings and items based on these basic configuration of objects. Program call with the same class after in the acquisition of sufficient configuration information, finish association analysis according to different rules in different methods. The analysis result is provided to the user for reference.

References

[1]http://www.infosecurity-us.com/view/11292/. cyber crimes cost organizations 38 million per year. 27 July 2010

[2]http://www.edu.cn. network black industry challenge network security 04 Jan 2011

[3]Litao Zhang. Security management model research based on system boundary analysis information. 2005

[4]Jinbo Chen. The study on the application of data mining in telecommunication CRM. 2006

[5]PangNing Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. Pearson Addison Wesley, 2006 200-238

[6]Michael J.A.Berry, Gordon S.Linoff. Data Mining Techniques. Wiley Publishing. 2011 245-290